



# 16 edycja konferencji SQLDay

13-15 maja 2024, WROCŁAW + ONLINE



partner platynowy



partner złoty



partner srebrny



# Use of Microsoft Fabric for data processing and analytics in practice

Full-day hands-on workshop

# Meet Your Instructors



**Estera Kot, PhD**  
Principal Product Manager  
Fabric DE&DS Product Group



**Bartłomiej Graczyk**  
Chief Technology Architect  
Data & Analytics



**Paweł Potasinski**  
Senior Program Manager  
Fabric Customer Advisory Team



**Jakub Wawrzyniak**  
Chief Technology Officer,  
Microsoft MVP



**Łukasz Grała**  
CEO, Futurologist,  
Microsoft MVP



**Maciej Rubczyński**  
Head of Development,  
Microsoft MVP



**Hubert Kobierzewski**  
BI Practice Lead,  
Data Community UG Leader



**Damian Widera**  
Data Solution Architect,  
Microsoft MVP



**Tomasz Libera**  
Data Architect,  
Microsoft MVP



Category	Description
Workshop Format	This is a hands-on workshop. You'll be actively involved in coding, implementing, and problem-solving.
What to Expect	Real-world scenarios and exercises. Direct application of concepts in live environments.
Participant Engagement	Please have your laptops and necessary tools ready. Prepare to code, collaborate, and share insights.
Hands-on	Please delve into the detailed descriptions provided for each exercise and step. Examine the attached screenshots carefully, and note the sequential visualization presented as 1), 2), 3). Review them thoroughly.
Support and Collaboration	Instructors and facilitators are here to guide you. Feel free to ask questions and help your fellow participants.
Outcome	By the end of this workshop, you'll have practical experience and new skills to apply directly to your projects.

# Agenda



- 08:30 – 09:15** (45 min) Introduction, Set Up and Overview of Fabric Data Platform (Presenter: Estera Kot & Pawel Potasinski)
- 09:15 – 10:35** (80 min) Exercise 1 - Ingest data with data pipelines and shortcuts (Presenter: Tomek Libera)
- 10:35 – 10:50** (15 min) Break
- 10:50 – 12:20** (90 min) Exercise 2 - Transform data using Notebooks and Spark clusters (Presenter: Jakub Wawrzyniak)
- 12:20 – 13:00** (40 min) Exercise 3 - Collaborate inside Notebooks and share Lakehouse. Use SQL Endpoint and SSMS (Presenter: Damian Widera)
- 13:00 – 14:00** (60 min) Lunch break
- 14:00 – 15:20** (80 min) Exercise 4 - Serve and consume data using Power BI and Data Science (Presenter: Hubert Kobierzewski)
- 15:20 – 16:20** (60 min) Exercise 5 - Latest Fabric Features (Presenter: Bartek Graczyk, Łukasz Grala)
- 16:20 – 16:30** (10 min) Break
- 16:30 – 17:00** (30 min) Exercise 6 – Open AI (Presenter: Maciej Rubczyński)
- 17:00 – 17:30** (30 min) Buffer, Recap and Extra exercises





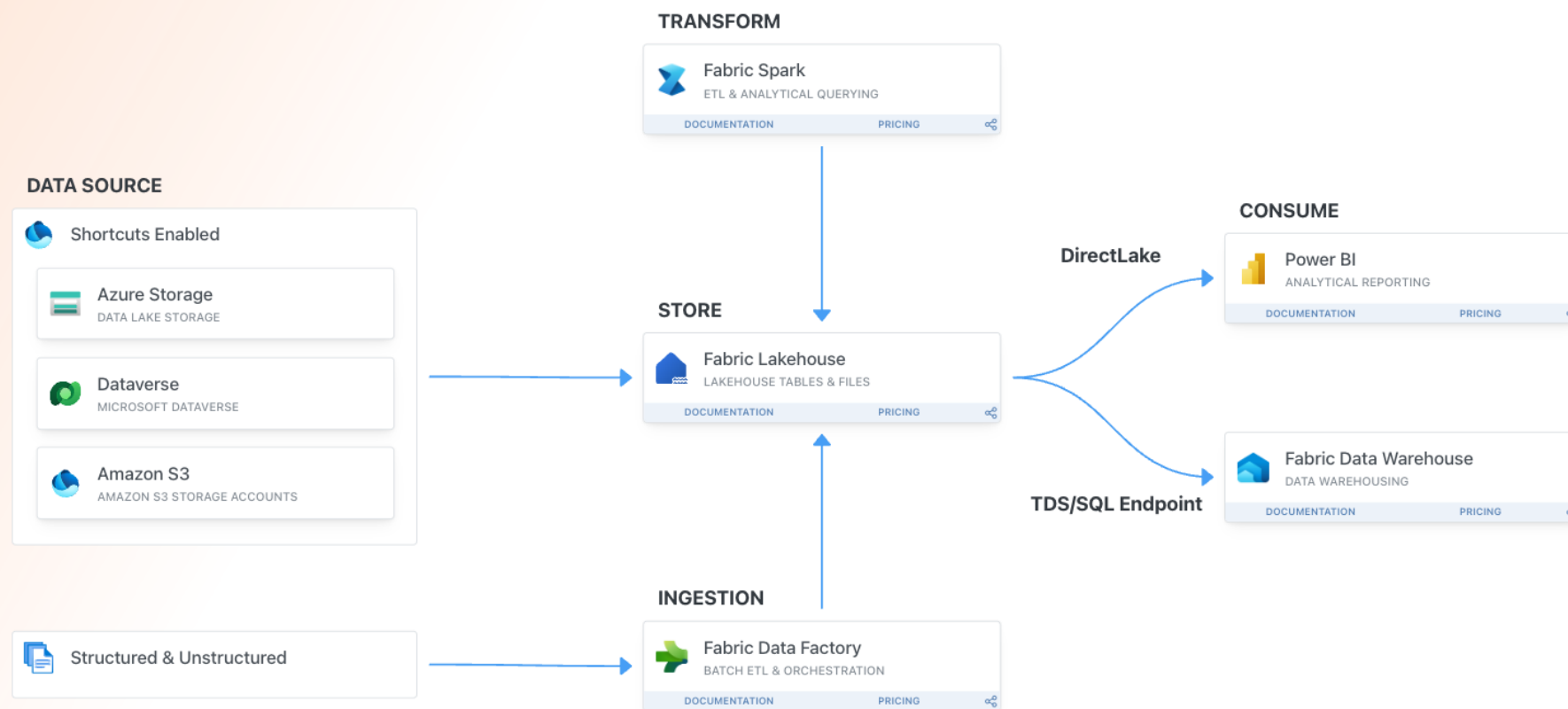
We have a WiFi for you



**Network:** SQLDay2024

**Password:** SQLDay%24

# High-level architecture



<https://aka.ms/sqlday-fabriclab>

**Enjoy, experience, network and learn!**



# Let's get to know each other!



## #1

Raise your hand if you have ever tried Microsoft Fabric before.

This could be in any context, just any prior exposure at all.

# Let's get to know each other!



## #2

Raise your hand if you have explored more than three contexts within Fabric, such as Power BI, Data Science, Data Factory, or Data Engineering.

# Let's get to know each other!



## #3

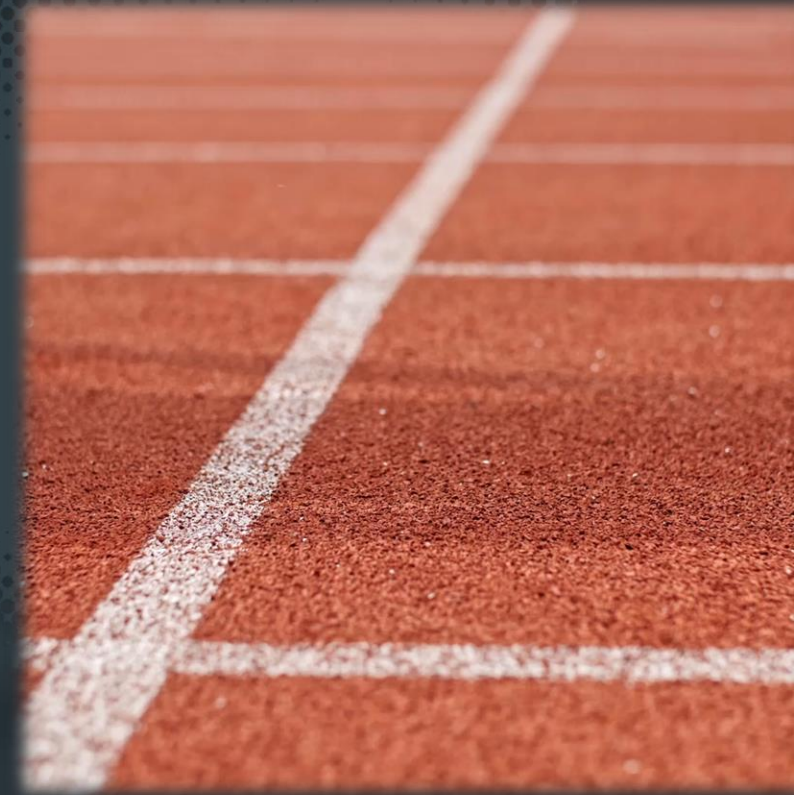
Raise your hand if you, or your company, have faced challenges with orchestrating and integrating different tools to ingest, process, and maximize the value from data.



# Introduction, Setup and Overview of Fabric Data Platform

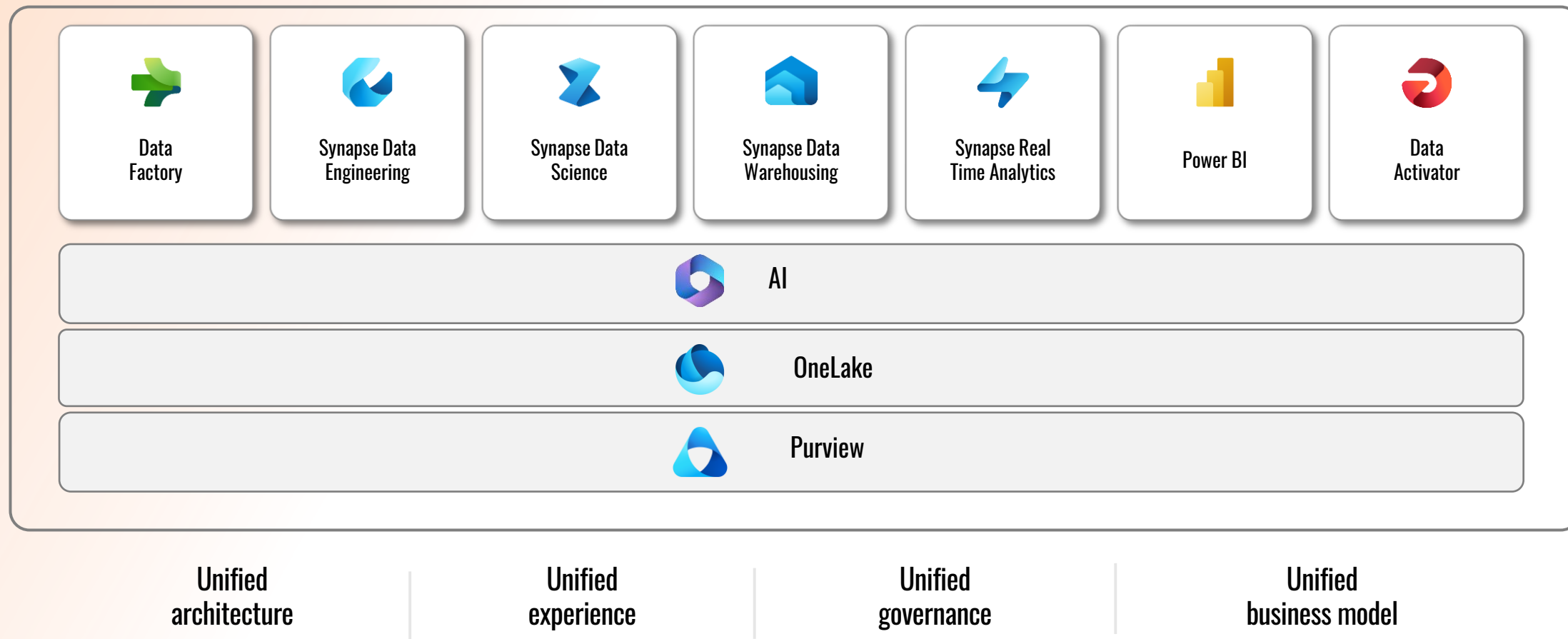


Setup and Overview of Fabric Data Platform





# The unified data platform for the era of AI





# The unified data platform for the era of AI



## Complete Analytics Platform

Everything, unified

---

SaaS-ified

---

Secured and governed

## Lake centric and open

OneLake

---

One Copy

---

Open at every tier

## Empower Every Business User

Familiar and intuitive

---

Built into Microsoft 365

---

Insight to action

## AI Powered

Copilot accelerated

---

ChatGPT on your data

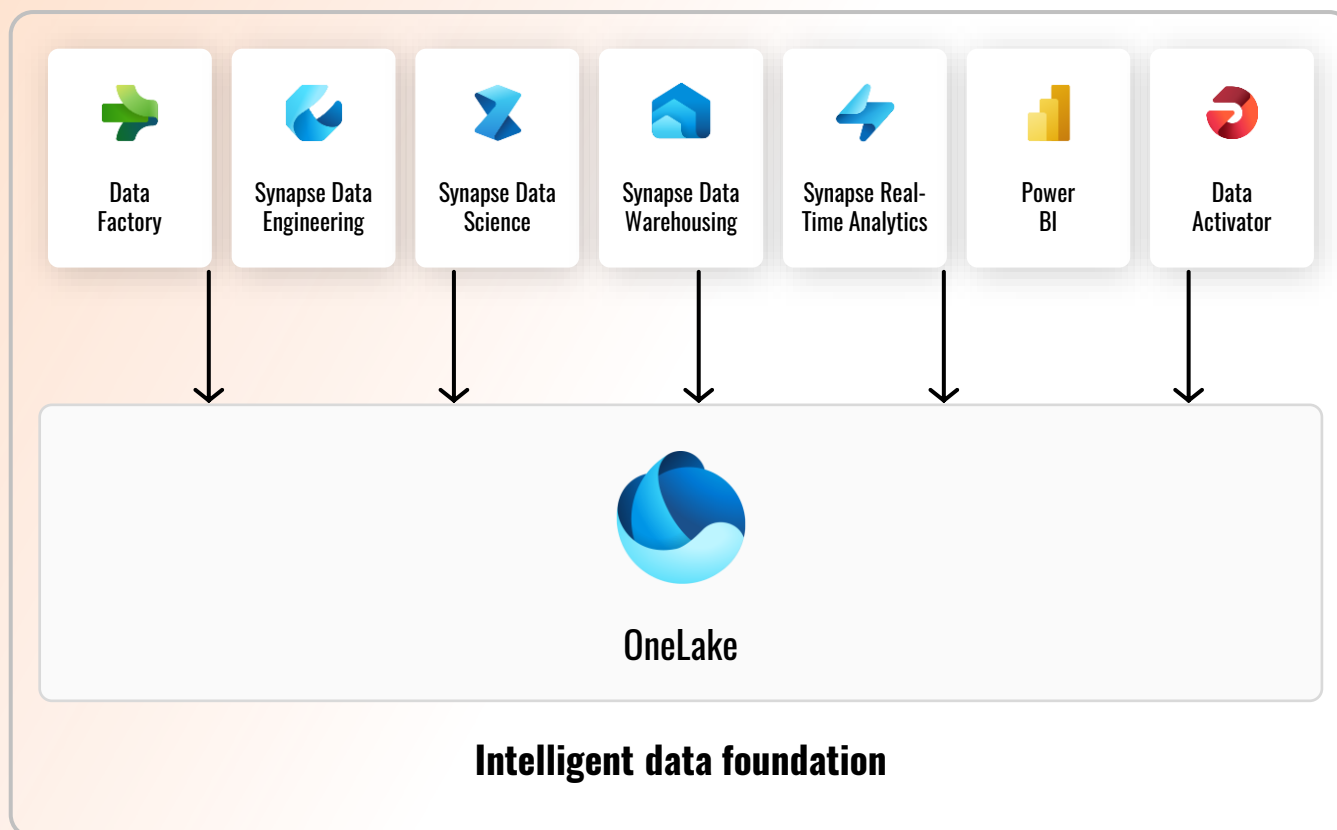
---

AI driven insights



# OneLake for all data

“The OneDrive for data”



A single SaaS lake for the whole organization

Provisioned automatically with the tenant

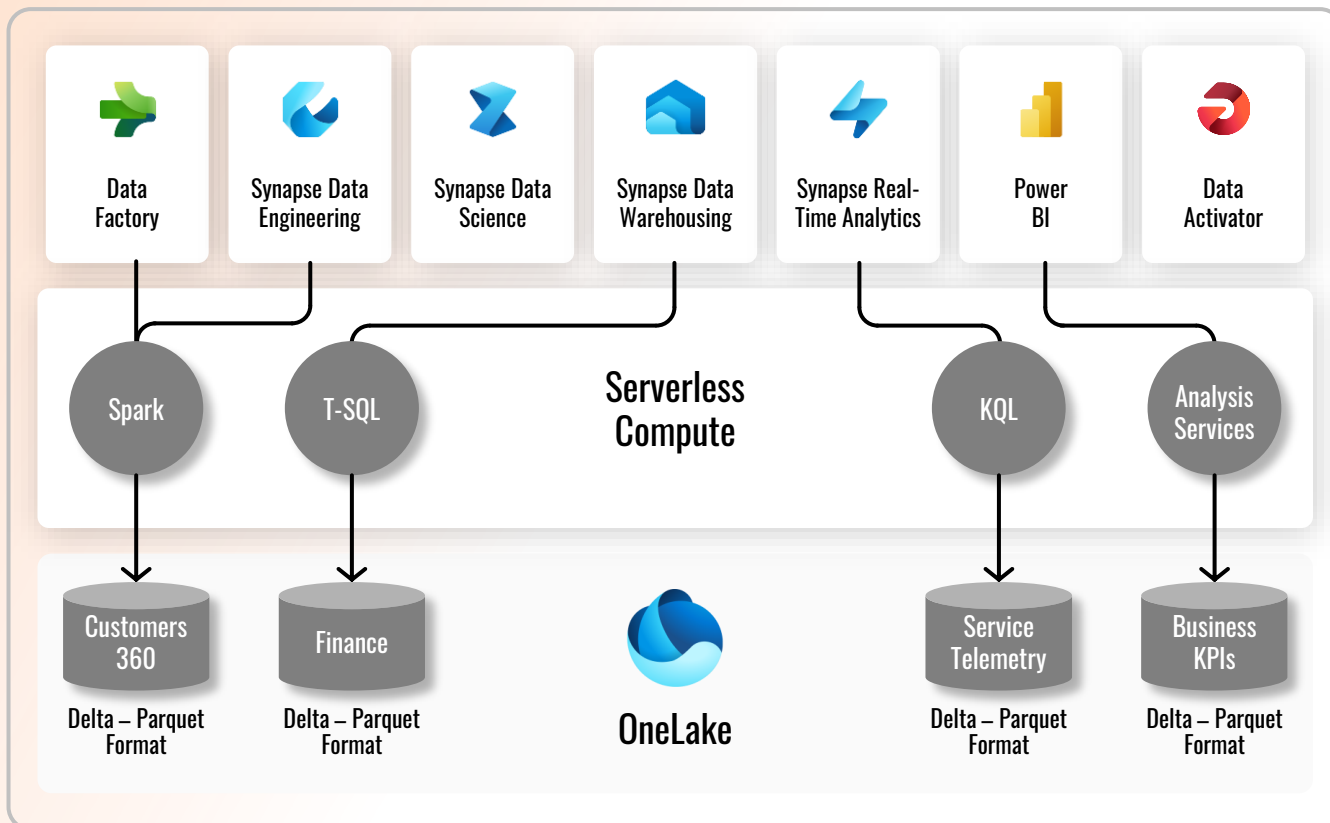
All workloads automatically store their data in the OneLake workspace folders

All the data is organized in an intuitive hierarchical namespace

The data in OneLake is automatically indexed for discovery, MIP labels, lineage, PII scans, sharing, governance, and compliance

# One Copy for all computes

## Real separation of compute and storage



All the compute engines store their data automatically in OneLake

The data is stored in a single common format

**Delta – Parquet**, an open standards format, is the storage format for all tabular data in Microsoft Fabric

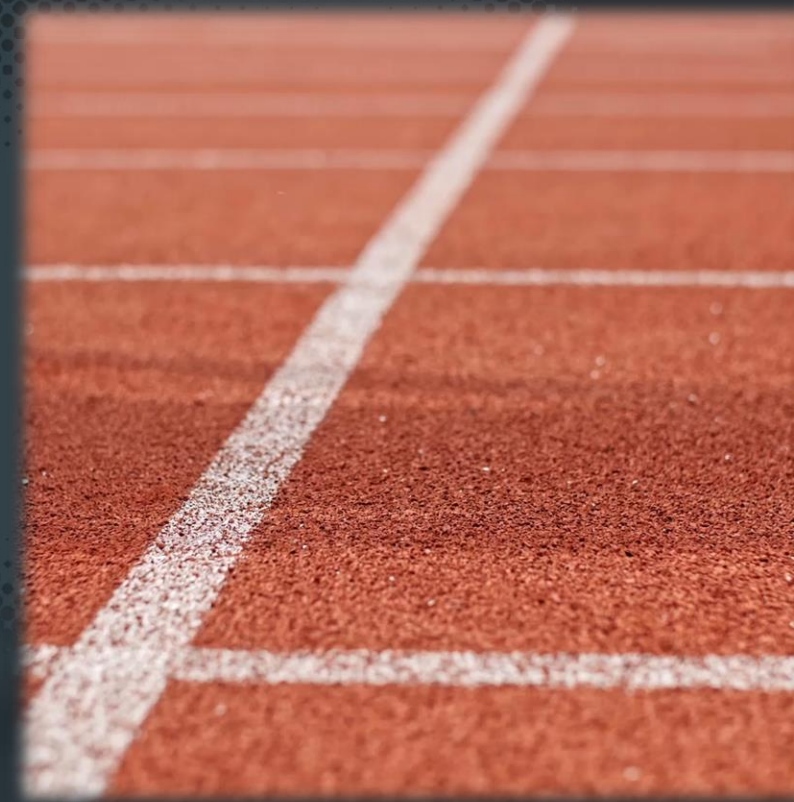
Once data is stored in the lake, it is directly accessible by all the engines without needing any import / export

All the compute engines have been fully optimized to work with Delta Parquet as their native format

Shared universal security model is enforced across all the engines

# Exercise 1

Ingest data with data pipelines and shortcuts



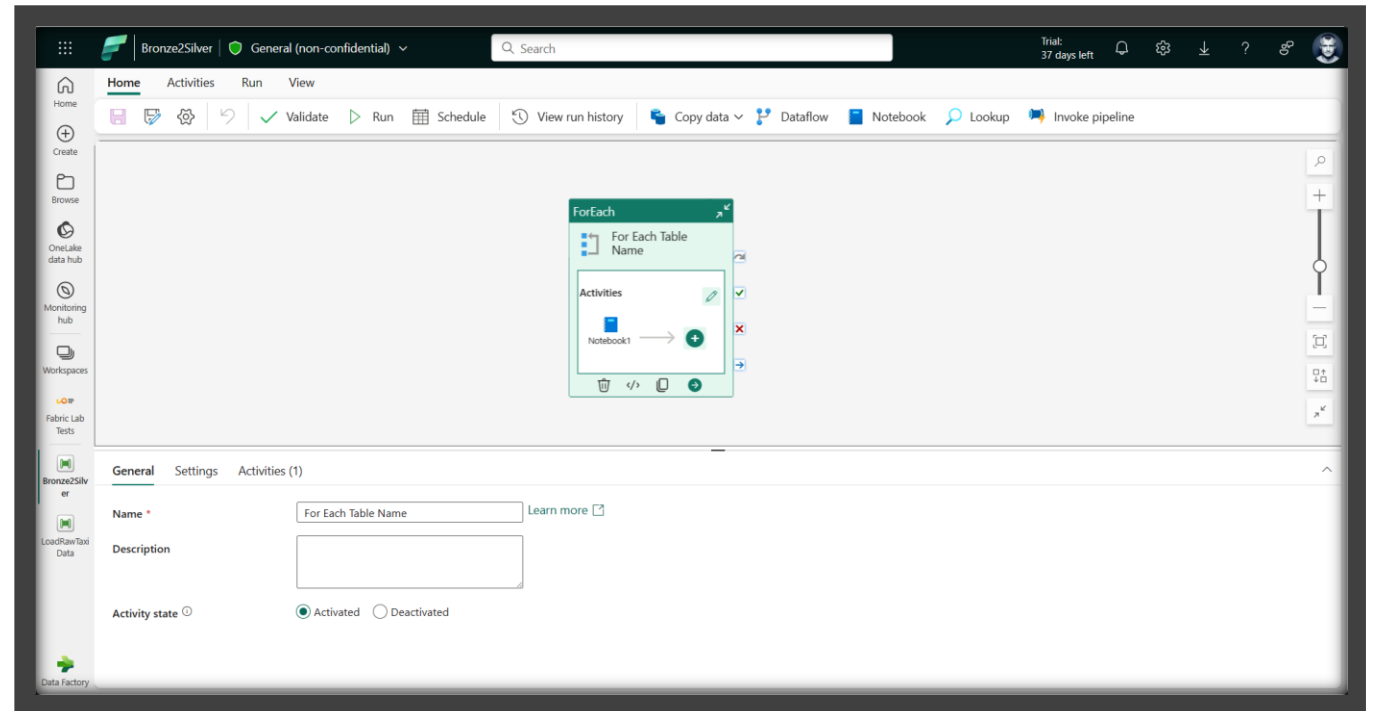


# Data Pipelines



**Data Pipelines enable powerful workflow capabilities at cloud-scale like building complex workflows, moving PB-size data, and defining sophisticated control flow pipelines.**







Data pipelines can be used to build complex ETL and data factory workflows that can perform a number of different tasks at scale. Additionally, control flow capabilities are built into pipelines so you can build workflow logic which provide loops and conditional.



# Activities

## Data transformation activities
















 Copy data ▾	<b>Copy data</b> from source to destination, including serialization/deserialization, compression/decompression, column mapping...
 Dataflow	Runs <b>Dataflow Gen2</b>
 Delete data	<b>Delete data</b> – files or folders from any supported stores
 Notebook	<b>Fabric Notebook</b> - runs Notebook created in Microsoft Fabric
 Stored procedure	Runs <b>stored procedure</b> in Fabric Data Warehouse or external (Azure SQL, SQL Server...)
 Script	Runs <b>SQL script</b> in Fabric Data Warehouse or external (Azure SQL, Snowflake...)

# Activities

## Control flow activities



 Get metadata	<b>Get metadata</b> - retrieve metadata of any data in a Data Factory or Synapse pipeline
 Lookup	<b>Lookup</b> - reads or look up a record/ table name/ value from any external source.
 Set variable	<b>Set variable</b> - sets the value of an existing variable
 If conditions	<b>If conditions</b> - the same functionality that an if statement in programming languages
 ForEach	<b>ForEach</b> - iterate over a collection and executes specified activities in a loop
 Web	<b>Web Activity</b> - calls a custom REST endpoint
 Invoke pipeline	<b>Invokes</b> another <b>pipeline</b>
 Switch	Implements a <b>switch</b> expression
 Filter	Apply a <b>filter</b> expression to an input array
 Wait	The pipeline <b>waits</b> for the specified time before continuing
 Until	Implements Do- <b>Until</b> loop
 Append variable	<b>Append variable</b> - adds a value to an existing array variable
 Fail	Cause pipeline execution to <b>fail</b> with a customized error message and error code



# Shortcuts

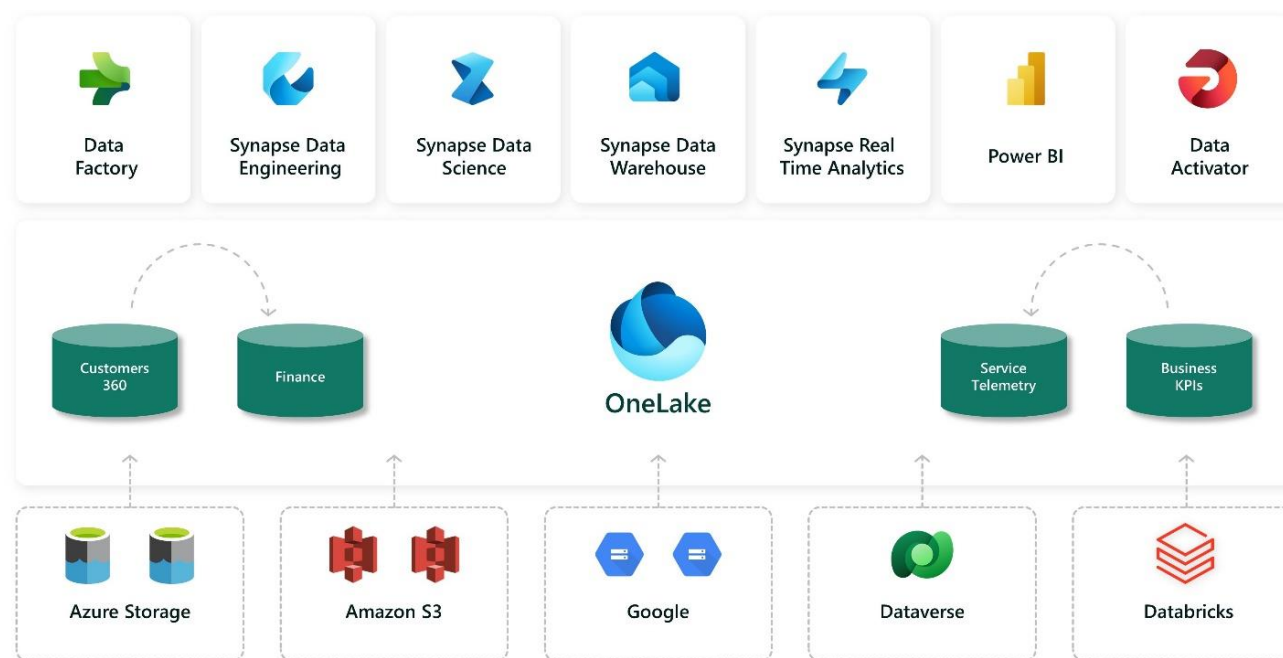


Shortcuts unify data without copying or moving existing data. This means that data can be used multiple times without data duplication.



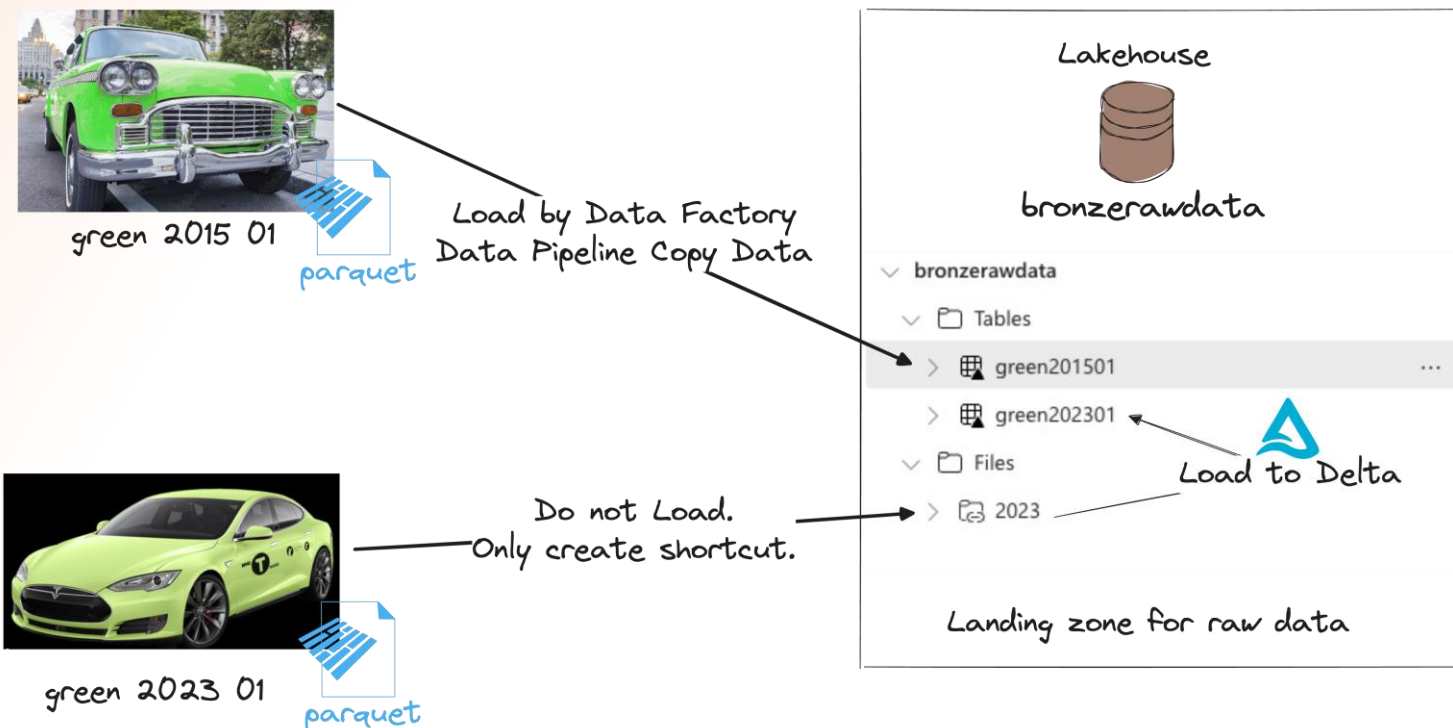
## Key Capabilities:

- Create shortcuts within Microsoft Fabric to consolidate data across artifacts or workspaces, without changing ownership of the data
- With shortcuts, data throughout OneLake can be composed together without any data movement
- Shortcuts also allow instant linking of data already existing in Azure and in other clouds, without any data duplication and movement, making OneLake the first multi-cloud data lake
- With support for industry standard APIs, OneLake data can be directly accessed by any application or service



# Exercise 1

Ingest data with data pipelines and shortcuts



Let's get your hands dirty with Fabric!

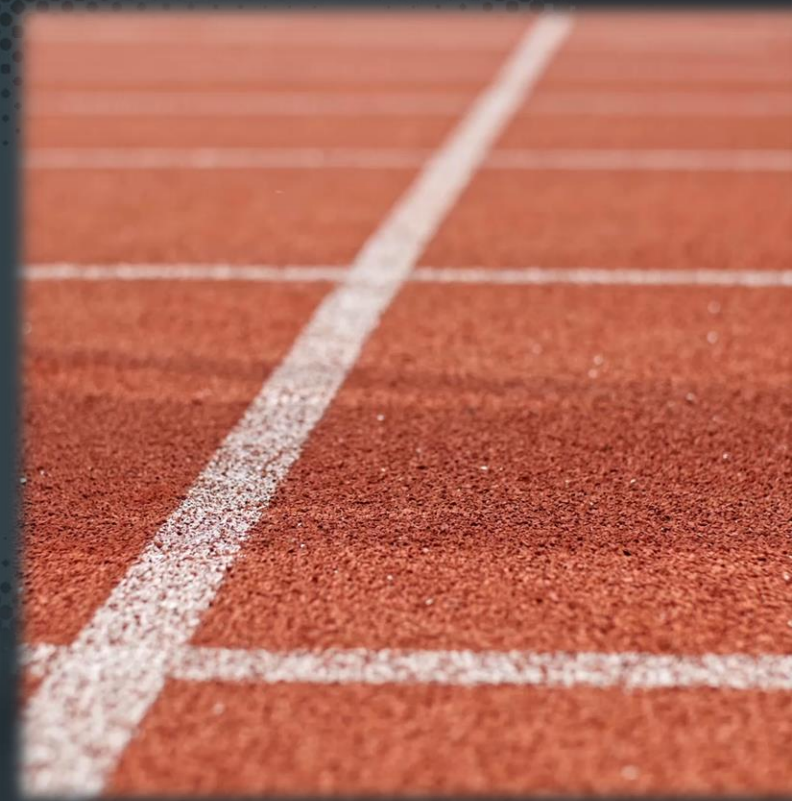


[aka.ms/sqlday-fabriclab](https://aka.ms/sqlday-fabriclab)



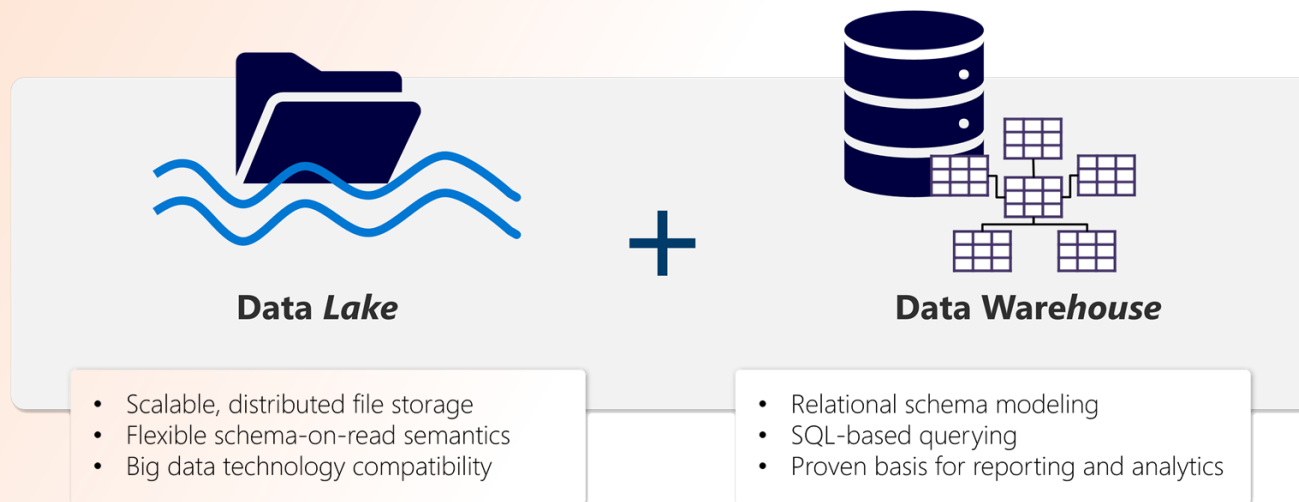
# Exercise 2

Transform data using Notebooks and Spark clusters



# Lakehouse concept

## Explore the Microsoft Fabric Lakehouse



Lakehouses use Spark and SQL engines to process large-scale data and support machine learning or predictive modeling analytics.

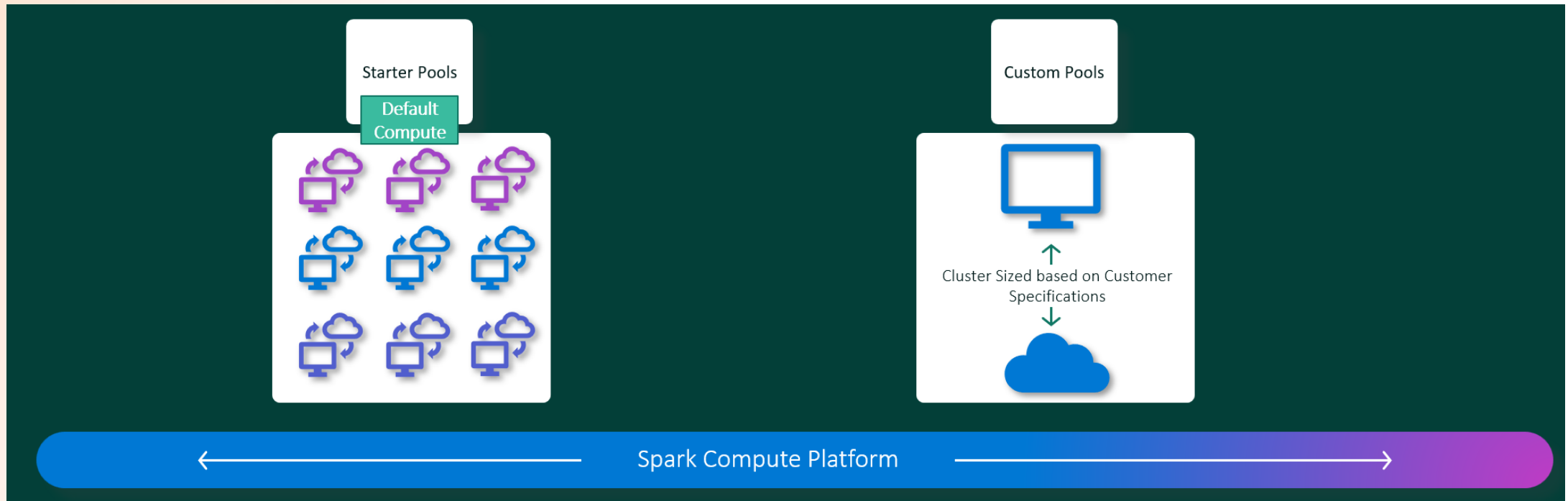
Lakehouse data is organized in a schema-on-read format, which means you define the schema as needed rather than having a predefined schema.

Lakehouses support **ACID** (Atomicity, Consistency, Isolation, Durability) transactions through Delta Lake formatted tables for data consistency and integrity.

Lakehouses are a single location for data engineers, data scientists, and data analysts to access and use data.

# Spark compute for Data Engineering

## Spark compute platform





# Spark compute for Data Engineering

## Starter pools



### Starter Pool Configuration

Node family	Memory optimized
Node Size	Medium
Min and Max Nodes	[1, 10]
Autoscale	On
Dynamic Allocation	On

# Data Engineering in Microsoft Fabric

## Microsoft Fabric Data Engineering items



Lakehouse



Notebook



Environment  
(Preview)



Spark Job  
Definition



Data pipeline



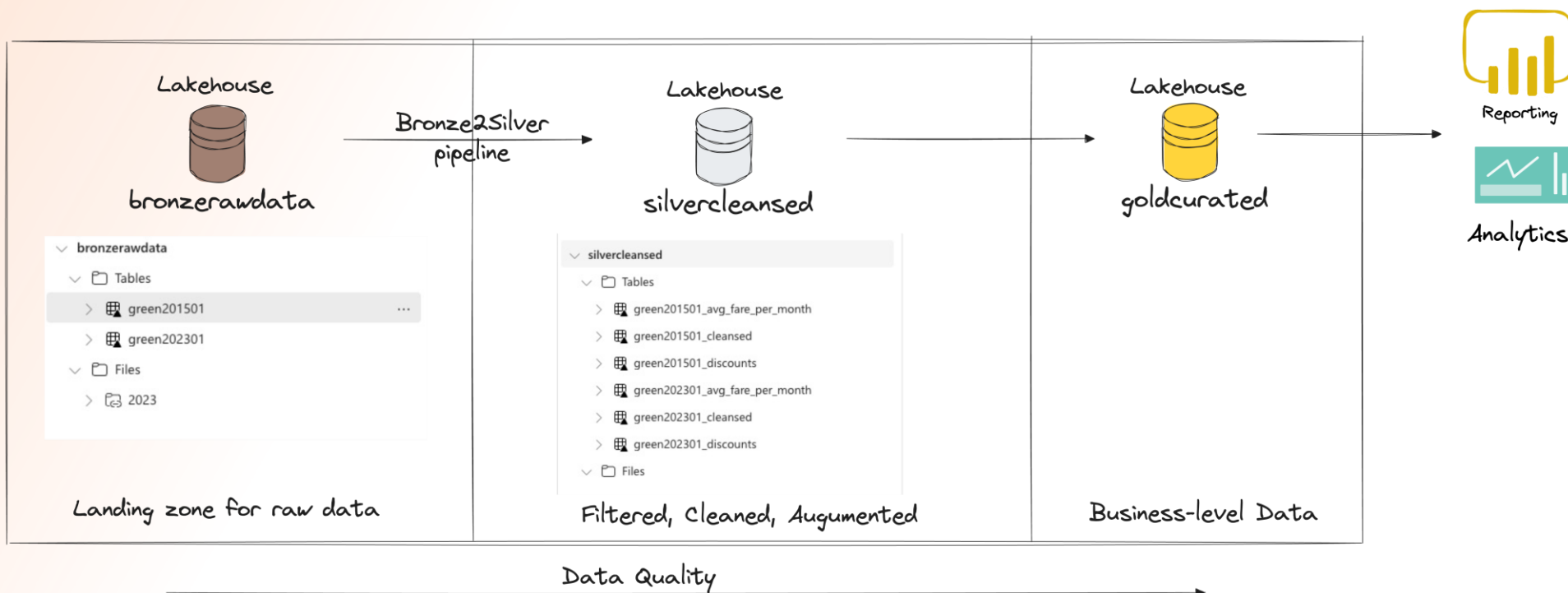
Import notebook



Use a sample

# Exercise 2

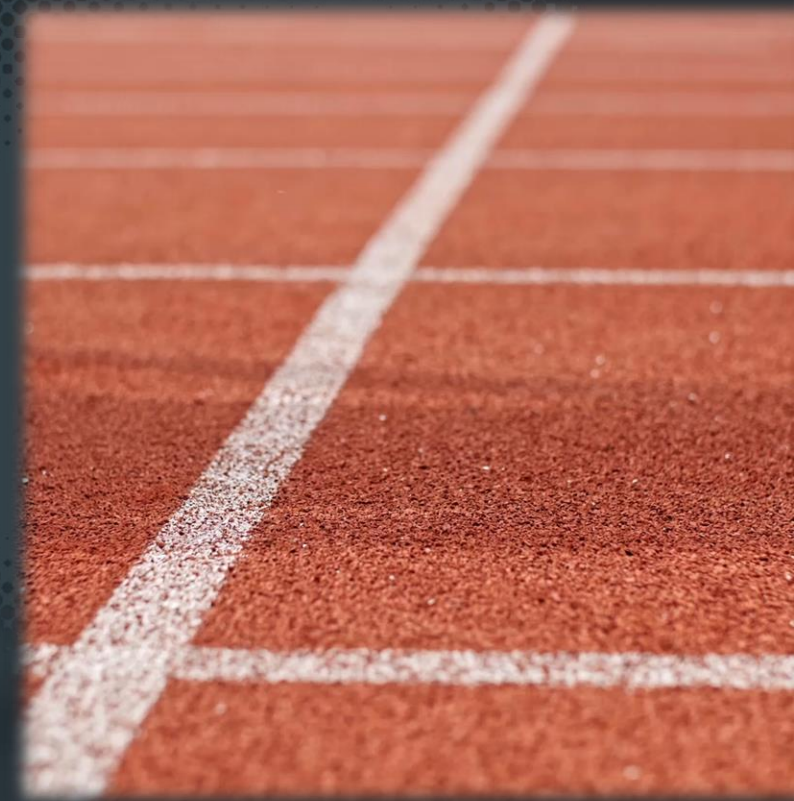
Transform data using Notebooks and Spark clusters





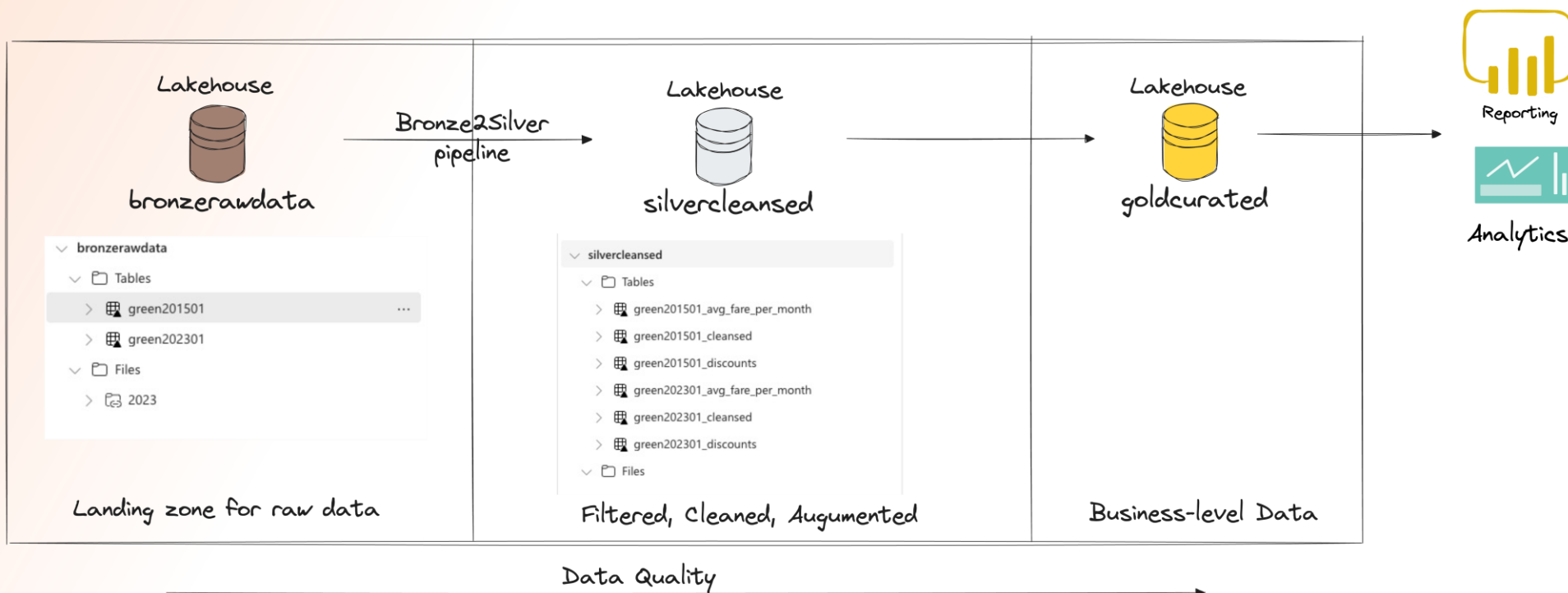
# Exercise 3

Collaborate inside Notebooks and share Lakehouse.  
Use SQL Endpoint and SSMS



# Exercise 3

Collaborate inside Notebooks and share Lakehouse. Use SQL Endpoint and SSMS



# Exercise 3

- Retrieve Lakehouse SQL analytics endpoint connection string
- Connect to Fabric SQL endpoint from SSMS
- Execute T-SQL queries on Lakehouse Delta Tables
- Share your Lakehouse
- Share your notebook for collaboration

# Get lakehouse's SQL analytics endpoint connection string



The screenshot illustrates a three-step process in the Power BI interface to obtain a SQL connection string from a Lakehouse:

- Step 1:** In the 'Lakehouse' view, the 'silvercleansed' Lakehouse is selected. The 'SQL analytics endpoint' option is highlighted, and the 'More options' menu is opened.
- Step 2:** The 'More options' menu is displayed, and the 'Copy SQL connection string' option is selected.
- Step 3:** A confirmation toast message appears, stating 'Copied to clipboard!'. It provides the connection string: `krrvjodmlo3ehogm2woebho6je-orx7llx...` and includes a 'Copy' button.



# Connect to Fabric SQL endpoint from SSMS

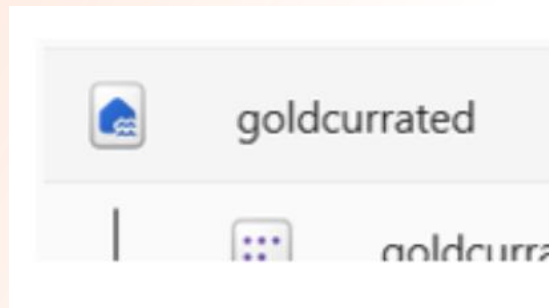


The screenshot shows the 'Connect to Server' dialog box with the following fields and values:

- Server type:** Database Engine
- Server name:** 4d3e5fsgalhweq2eri.datawarehouse.fabric.microsoft.com
- Authentication:** Microsoft Entra Password (highlighted with a green box)
- User name:** BB001@fabricconf.onmicrosoft.com
- Password:** (empty field)
- ☐ Remember password

Buttons at the bottom: Connect, Cancel, Help, Options >>

# Share your Lakehouse



### Grant people access

goldcurrated

People you share this Lakehouse with can open it and its SQL endpoint and read the default dataset. To allow them to read directly in the Lakehouse, grant additional permissions.

1

DE001

Enter a name or email add

2

#### Additional permissions

☐ Read all SQL endpoint data ⓘ

☐ Read all Apache Spark ⓘ

☐ Build reports on the default dataset

3

#### Notification Options

☒ Notify recipients by email

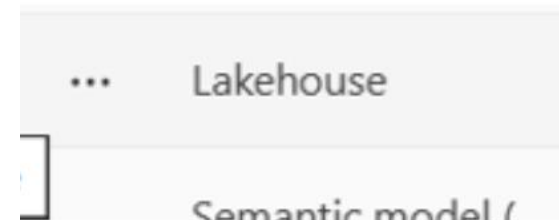
Add a message (optional)

4

Grant

Back

Depending on which additional permissions you select, recipients will have different access to the SQL endpoint, default dataset, and data in the lakehouse. For details, view lakehouse permissions documentation.



# Share your no



1

## Select permissions

1 Test Widgets

People who can view this Notebook



People in your organization



People with existing access



Specific people

## Additional permissions

Authorized users can view this Notebook by default.  
Select additional permissions.



Share



Edit



Run



Notebook artifact Sharing

Apply

Back

Language

PySpark (Python)



Comments

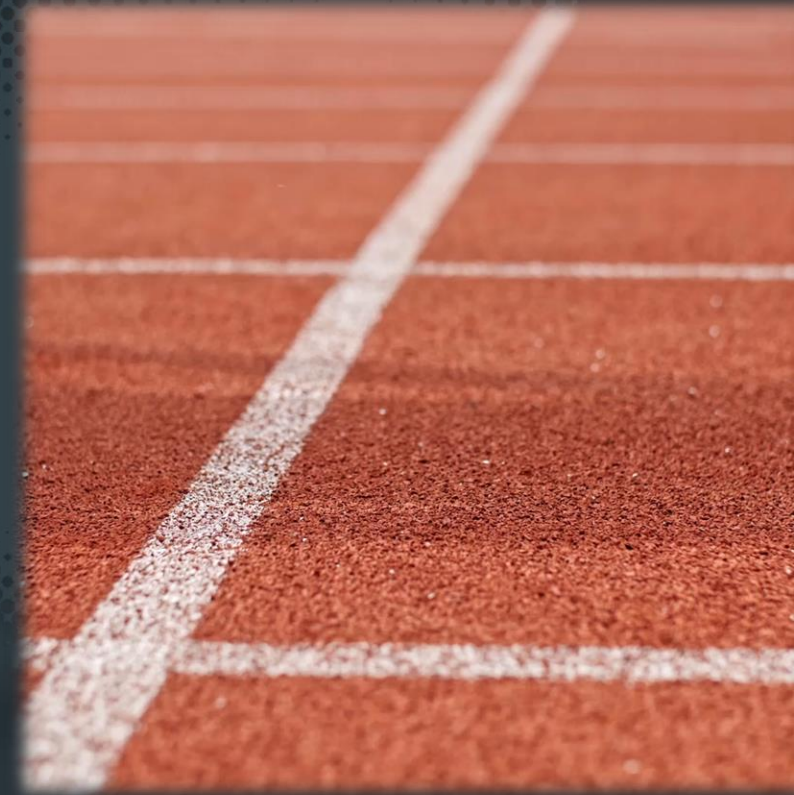


Share



# Exercise 4

Serve and consume data using Power BI and Data Science





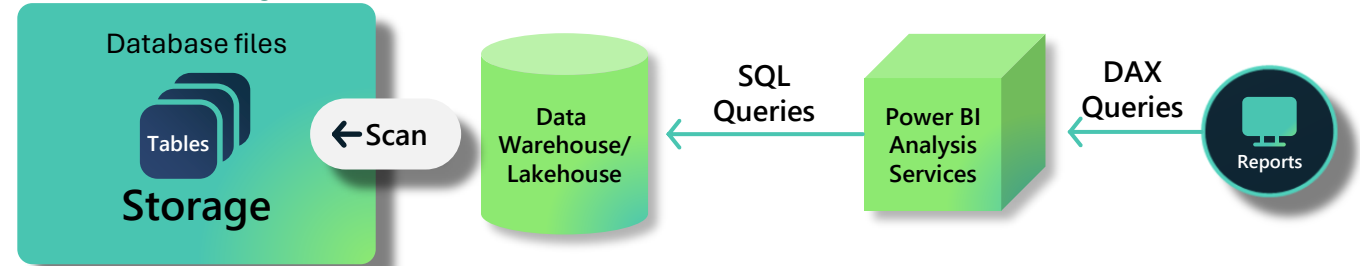
# Power BI | Direct Lake mode



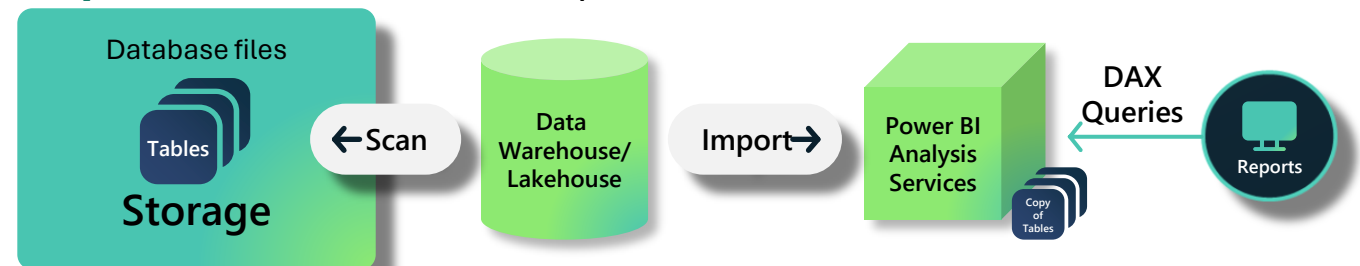
**Direct Lake** is a fast-path to load the data from the lake straight into the Power BI engine, ready for analysis.

Direct Lake is based on loading parquet-formatted files directly from a data lake without having to query a Lakehouse endpoint, and without having to import or duplicate data into a Power BI dataset.

**Direct Query mode.** Slow, but real time



**Import mode.** Latent and duplicative, but fast



**Direct Lake mode.** Fast and real time



# How Direct Lake Mode is different?

- Currently supports only one data source
- When querying, data needs to be loaded up (paged) into memory
- But not whole tables... only needed columns
- Based on Delta table format – parquet files, best with V-Order
- You need to start with Lakehouse or Warehouse
- It may fallback and query underlying Warehouse or Lakehouse (too many rows!)
- Larger tables require “framing” management to keep report results consistent

# Direct Lake Mode limitations

- String data type is limited to 4000 characters
- No relationships based on DateTime data type
- Source data based only on tables (no views)
- No calculated columns
- No calculated tables
- Development only with Fabric web-based editor or via XMLA endpoint (currently Tabular Editor supports it)
- Not supported on PBI Pro, PBI PPU, neither on PBI Embedded A/EM

# Why is V-Order better?

- Technology present in Microsoft since 2009
- Proprietary optimization technique for writing data in parquet files
- Same algorithms as in PBI Import Mode
- Once data is saved in files, it can be read by any tool that works with Delta tables



# Data science in Microsoft Fabric

End-to-end data science for predictive business insights



## Data Centric

- Easy and secure access to lake-centric data
- Open Delta Lake support promotes reproducibility
- Native integration with data infrastructure



## Developer friendly

- SaaS experiences with quick setup
- Code authoring experiences in Notebooks and IDE
- VS Code integration



## Rich ML tools

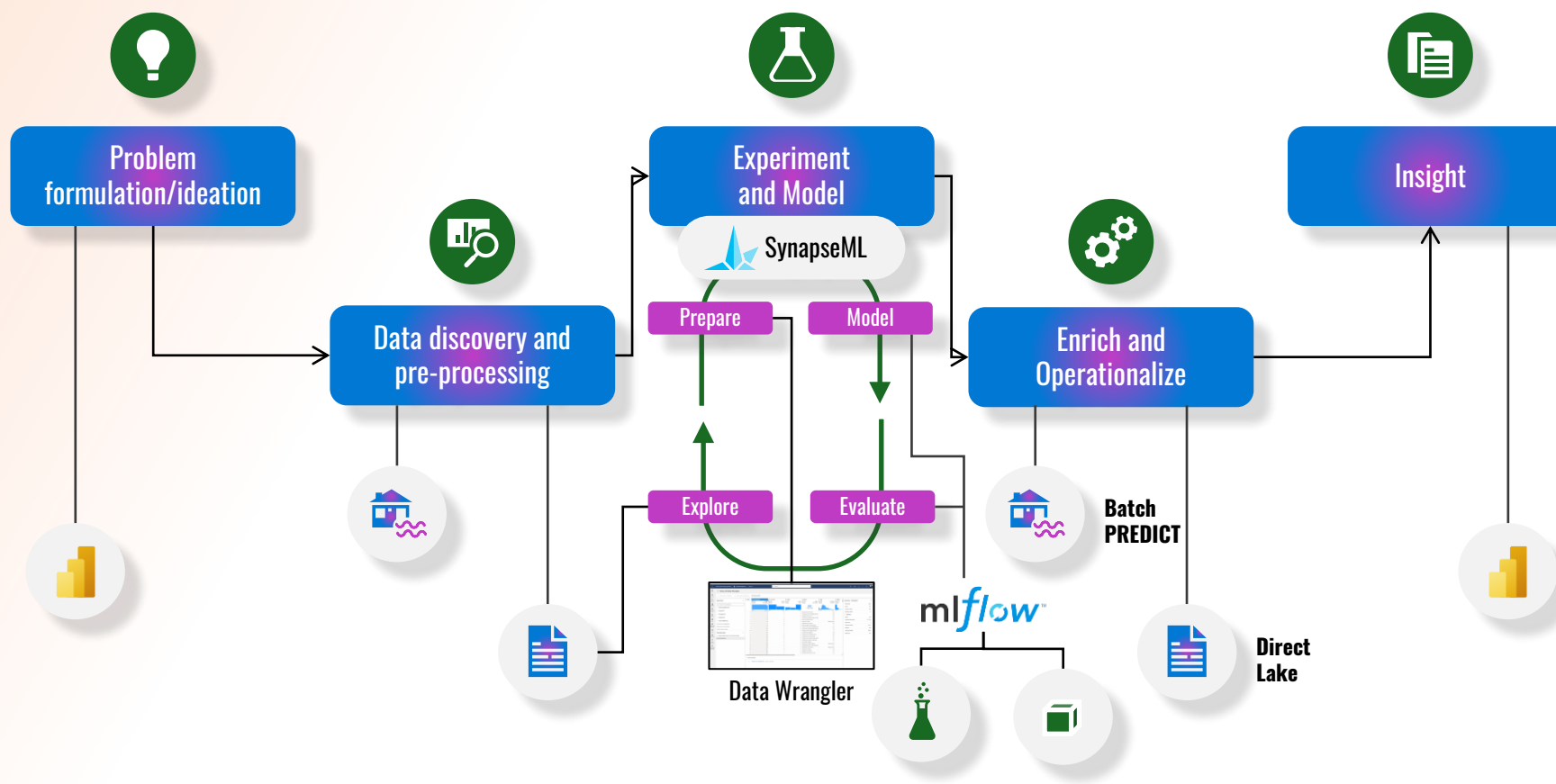
- Supports MLFlow model and experiment management
- Built-in, scalable ML tools with SynapseML
- Direct Lake mode for serving predictions to BI reports



## Promotes collaboration

- Unified platform for all analytics roles incl. data scientists
- Secure and easy sharing of data, code, models and experiments

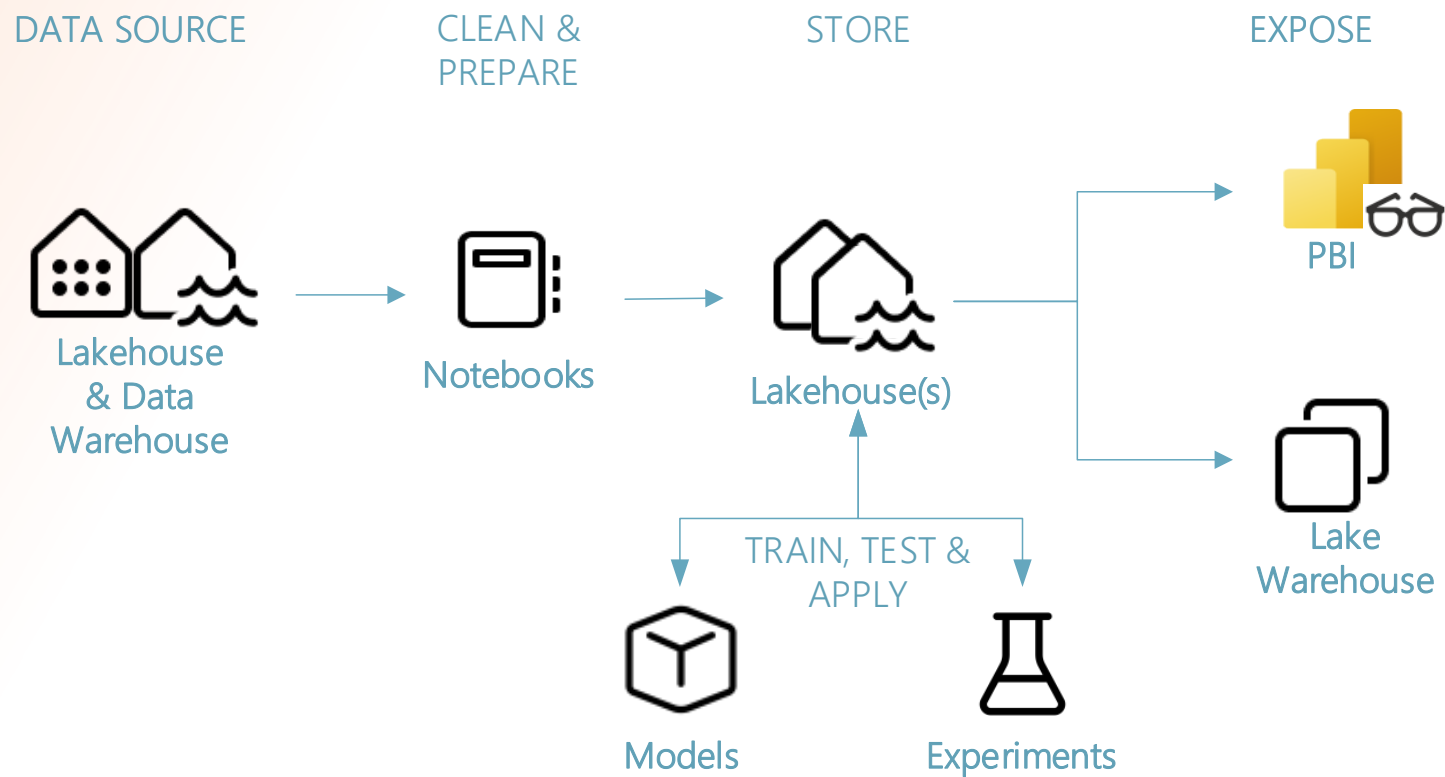
# The Fabric Data Science experience



# Exercise 4

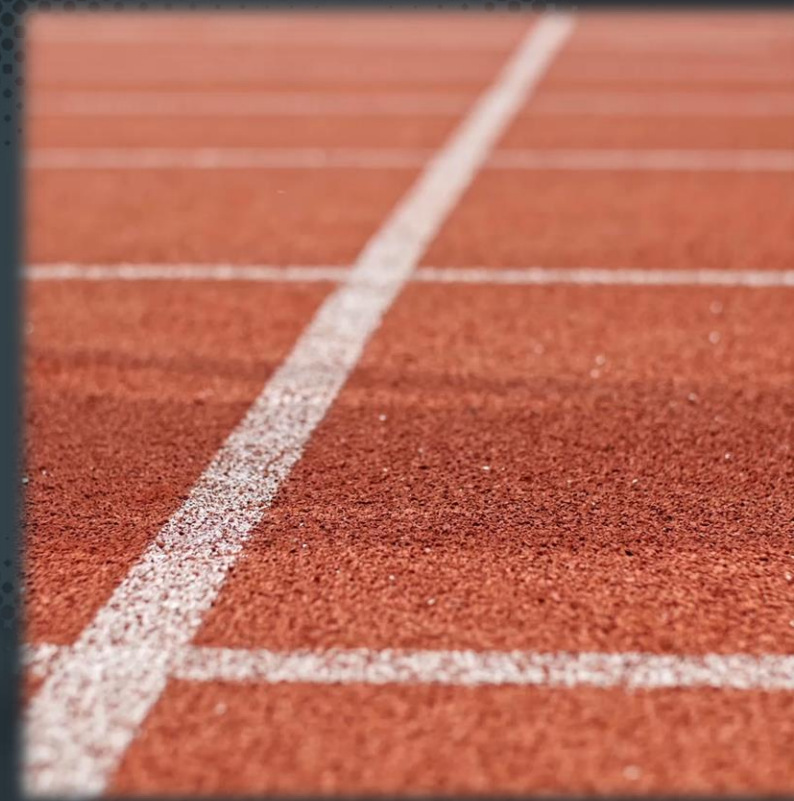


Serve and consume data using Power BI and Data Science



# Exercise 5

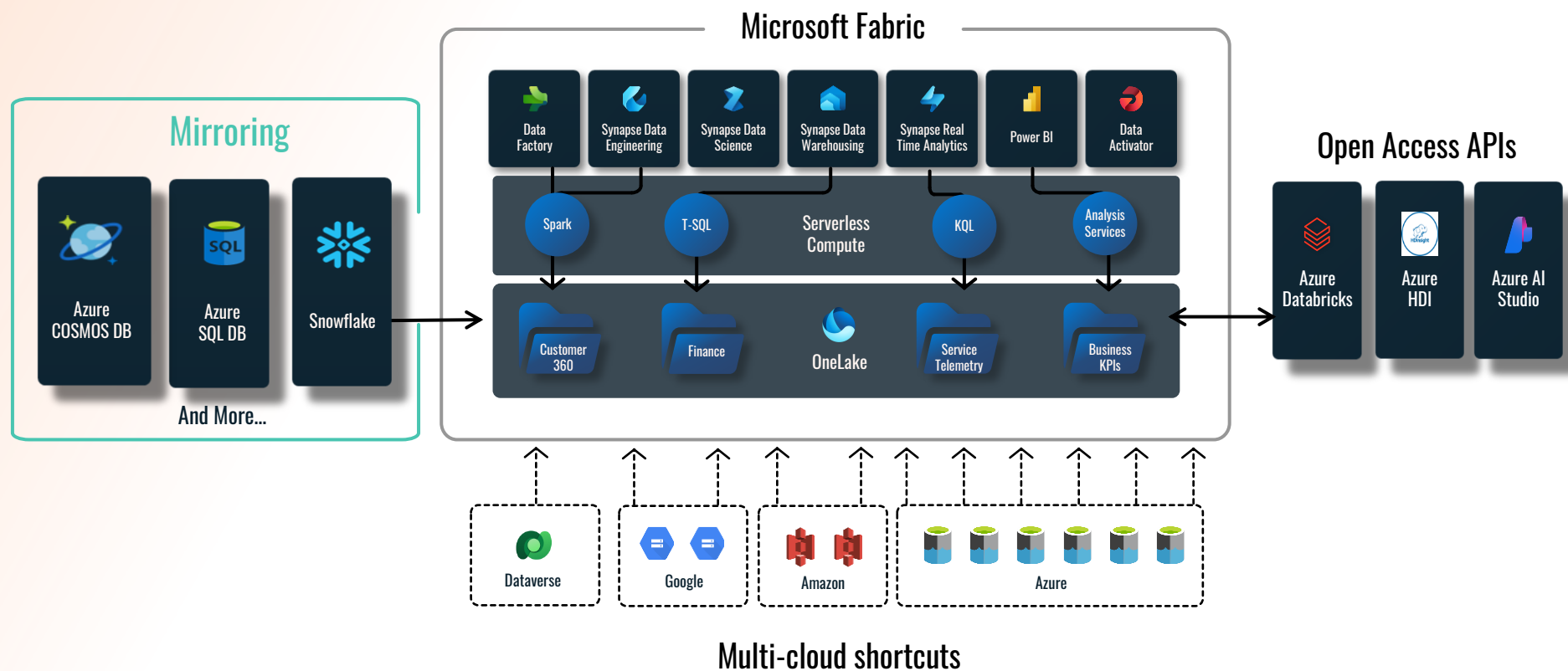
Latest Fabric Features





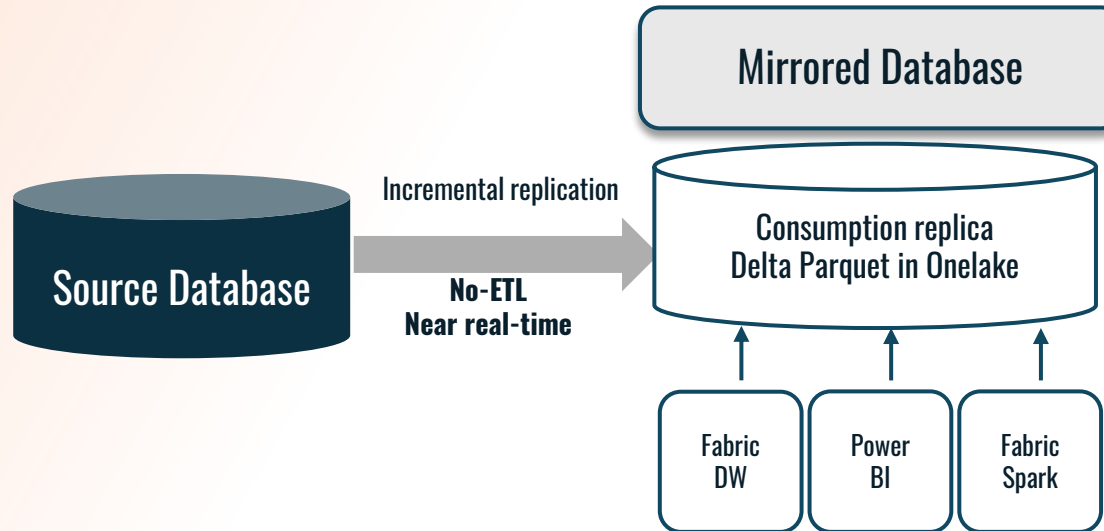
# Creating Data Gravity in OneLake

## Connect your entire data estate



# Mirroring in Microsoft Fabric

## Simplify near real-time analytics



Fabric Mirroring enables adding existing databases and data warehouses to Fabric without any ETL.

Data is replicated into OneLake in Delta format in near real-time

All of the Fabric analytics and AI experiences instantly work with the Mirrored Database

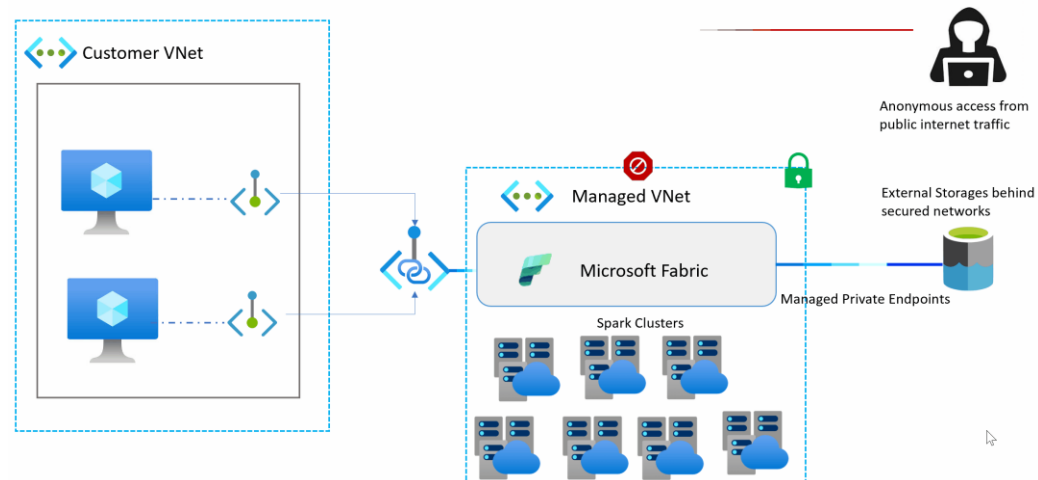
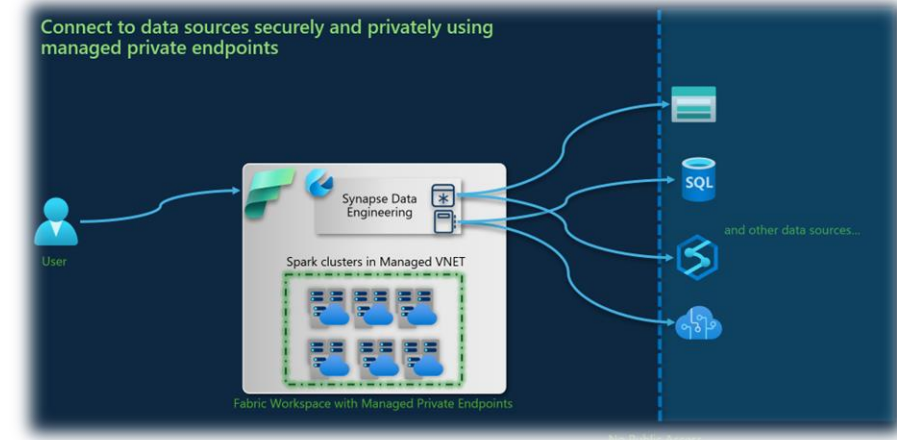
Your data is in an open format enabling limitless opportunities for data liberation

# Managed private endpoints in Microsoft Fabric

## Connect to data sources securely



- With a managed virtual network you get complete network isolation for the Spark clusters running your Spark
- You don't need to create a subnet for the Spark clusters based on peak load, as this is managed for you by Microsoft Fabric.
- A managed virtual network for your workspace, along with managed private endpoints, allows you to access data sources that are behind firewalls or otherwise blocked from public access.

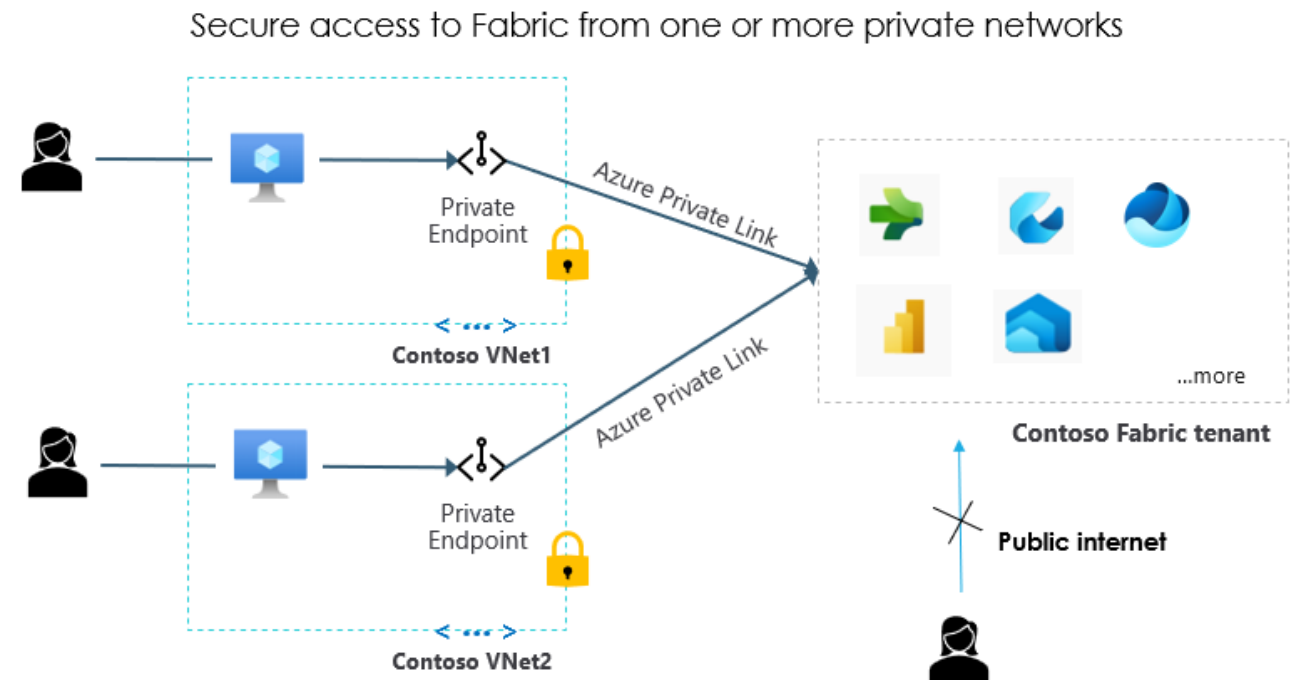


# Private links in Microsoft Fabric

## Connect to data sources securely



- Restrict traffic from the internet to Fabric and route it through the Microsoft backbone network
- Ensure only authorized client machines can access Fabric
- Comply with regulatory and compliance requirements that mandate private access to your data and analytics services.



<https://aka.ms/fabricsecuritywhitepaper>

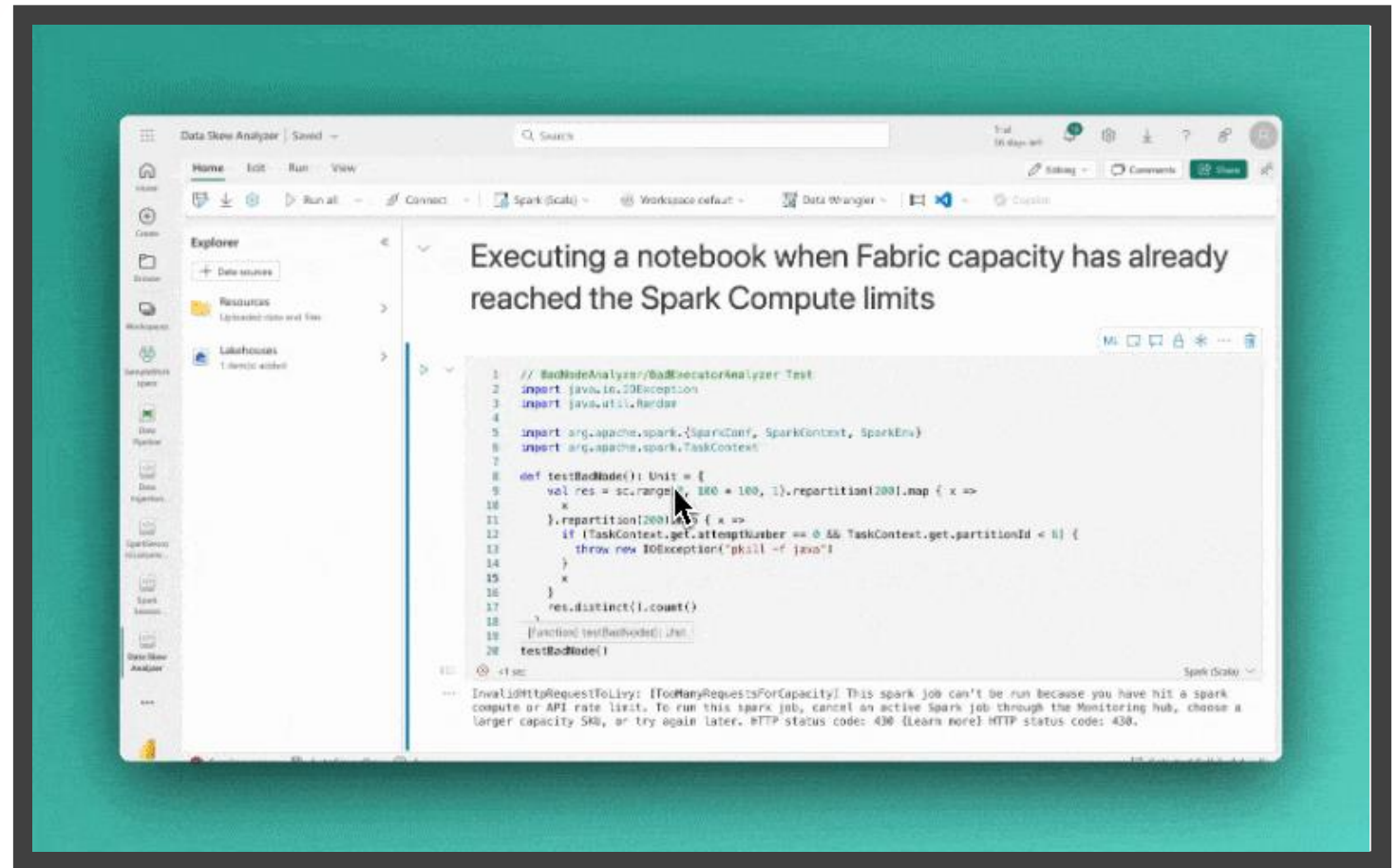


# Job Queueing for Notebook



Job Queueing eliminates manual retries and improve the user experience for users who run notebook jobs on Microsoft Fabric.

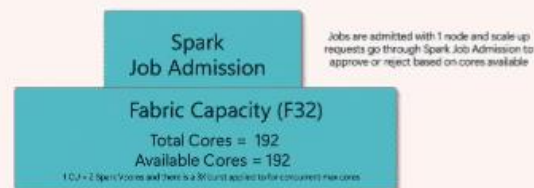
Notebook jobs that are triggered by pipelines or job scheduler will be added to a queue and will be retried automatically when the capacity frees up.



# Optimistic Job Admission for Spark

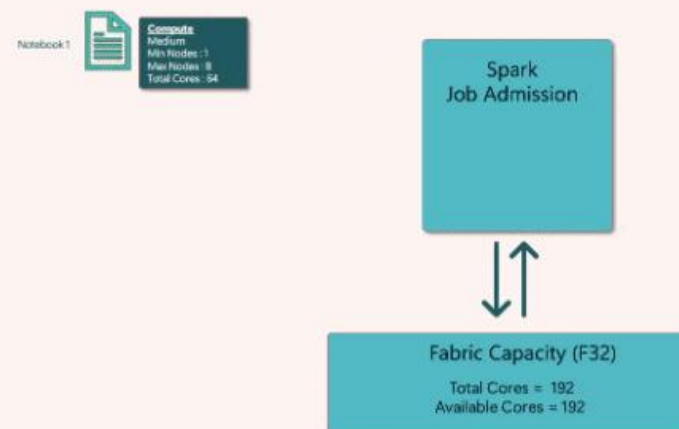
## Job Concurrency (Without Optimistic Job Admission)

### Job Admission (Optimistic Approach)



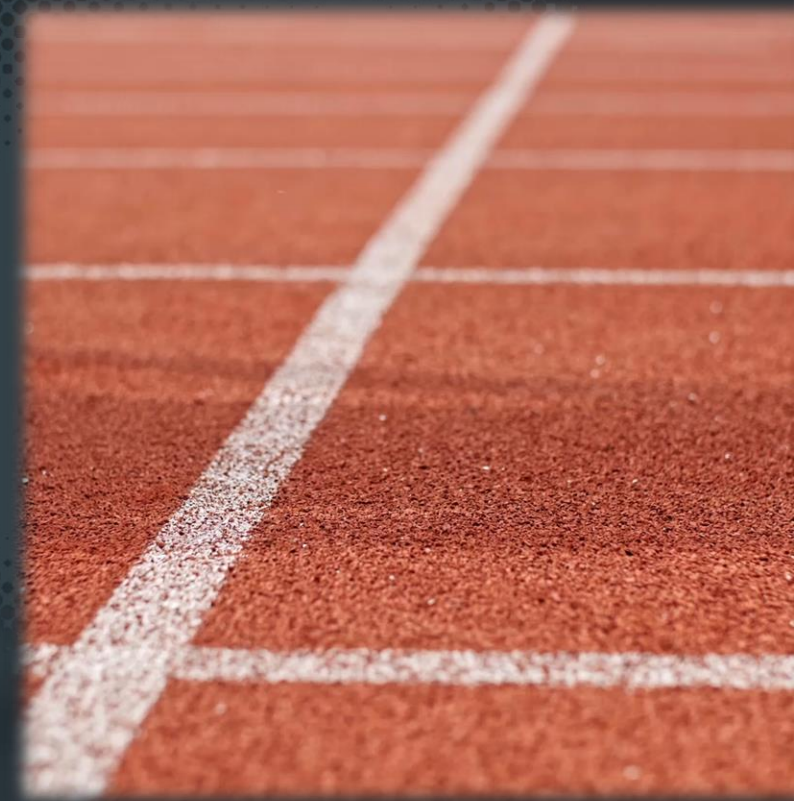
- By enabling autoscale, F32 capacity could support running 24 concurrent jobs (192 Total Cores/8 Core per job)
- Job scale up is approved/ rejected based on cores available in a fair manner
- Scale up or new job admission exceeding the available cores are queued or throttled

### Job Scale Up Flow



# Exercise 6

Prebuilt Azure AI Services and Azure OpenAI





# Prebuilt AI models in Fabric

- Fabric seamlessly integrates with Azure AI services, allowing you to enrich your data with prebuilt AI models without any prerequisite.
- You can utilize your Fabric authentication to access AI services.
- All usage are billed against your Fabric capacity
- Currently in public preview with limited AI services available.



**Azure  
OpenAI**



**Text  
Analytics**



**AI Translator**



# Prebuilt AI models in Fabric



**AI Translator**

- Translate: Translates text
- Transliterate



**Text Analytics**

- Language detection
- Sentiment analysis
- Key phrase extraction
- Personally Identifiable Information(PII) entity recognition
- Named entity recognition
- Entity linking

# Prebuilt AI models in Fabric



- Text generation
- Code generation
- Embeddings
- Classification
- All previous tasks: sentiment analysis, translation, language detection etc.



**Azure  
OpenAI**

- GPT-35-turbo
- gpt-4 family
- text-embedding-ada-002 (version 2)
- text-davinci-003,
- code-cushman-002

# Workshop recap



# Microsoft Fabric Resources



## Community Call to Action

---

- ✓ Try Microsoft Fabric for free: <https://aka.ms/try-fabric>
- ✓ Join the Fabric community: <https://aka.ms/fabriccommunity>
- ✓ Share and vote for ideas to improve Fabric: <https://aka.ms/fabricideas>
- ✓ Read and comment our blog: <https://aka.ms/fabricblog>

## Learn More About Microsoft Fabric

---

- Product website: <https://aka.ms/microsoft-fabric>
- Documentation: <https://aka.ms/fabric-docs>
- Fabric e-book: <https://aka.ms/fabric-get-started-ebook>
- Microsoft Learn: <https://aka.ms/learn-fabric>
- End-to-end scenario tutorials: <https://aka.ms/fabric-tutorials>
- Fabric Notes: <https://aka.ms/fabric-notes>